

EVALUATING OPEN AI TOOLS O1 AND O3-MINI IN SOLVING HIGH SCHOOL PROBLEMS FROM SERBIAN NATIONAL MATHEMATICS COMPETITION

Aleksandar Milenković, Nemanja Vučićević, Marina Svičević

Abstract. Artificial intelligence (AI) tools are already being widely used in mathematics education in various ways and for different purposes. One important direction of their application lies in using AI tools for solving complex, competition-level mathematical problems. In this study, we contribute to this field by analyzing and evaluating the solutions to problems that high school students solved at the National Mathematics Competition in the Republic of Serbia, which were generated by the OpenAI tools o1 and o3-mini. The results indicate that not all solutions produced by these two tools are accurate and fully correct. However, based on the evaluated solutions, it can be concluded that, in most cases, the performance of these two tools would qualify them for one of the top three prizes in competition with the students. Experienced students, mathematics competitors who qualified for the national competition – could benefit from reviewing the solutions to problems they were unable to solve independently, provided by o1 and o3-mini, as long as they approach these solutions critically and carefully analyze each aspect of the reasoning process.

MathEduc Subject Classification: P84

AMS Subject Classification: 97P80

Key words and phrases: National mathematics competition; high school mathematics problems; o1 tool, o3-mini tool, performance of AI tools.

1. Introduction

Artificial intelligence has found applications in various spheres of society, including education. Given the specific nature of mathematics as a discipline, as well as the particularities of mathematics instruction, there have been ongoing efforts to incorporate AI tools into the teaching and learning process of mathematics. These tools can serve different roles, with a special focus on supporting students in the process of solving mathematical problems.

Mathematics competitions have long been recognized worldwide as extremely beneficial for students, as participation itself enables them to broaden and deepen their knowledge while developing various competencies and skills. In recent years, researchers in the fields of mathematics education and artificial intelligence have been exploring the capabilities of existing AI tools and developing specialized models with the aim of successfully solving complex mathematical problems that appear in national and international competitions for high school students.

Considering that this topic has so far been insufficiently explored in research conducted in the Republic of Serbia, we aimed to examine the extent to which

two recently released and advanced OpenAI tools, o1 and o3-mini, are capable of successfully solving problems that students tackled at the National Mathematics Competition. Our goal was to shed light on whether competition participants can rely on these tools during their preparation and to what extent they can trust the solutions generated by these two AI models.

Theoretical background

Mathematical competitions

In numerous studies, authors emphasize that mathematical competitions play a significant role in supporting students with above-average achievements in mathematics [2]. Additionally, competitions serve as a means of identifying mathematically gifted students [12, 20]. Mathematical competitions provide students with the opportunity to come together, socialize, and measure their mathematical knowledge against their peers. Moreover, these competitions present a challenge not only for students but also for their mathematics teachers [20].

Beyond increasing interest in deeper engagement with and learning mathematics, preparing for mathematical competitions allows students to work on mathematical problems that go beyond the content covered in standard mathematics curricula [9]. At the same time, students tackle non-standard problems, fostering their critical and logical thinking skills [17]. During preparation, students invest effort in solving complex mathematical problems, thereby expanding and deepening their mathematical knowledge [9]. Interestingly, competition participants often go on to achieve success in scientific fields. Kenderov explains this phenomenon by pointing out that success in both mathematical competitions and scientific endeavors requires not only strong knowledge but also advanced intellectual capacities and a propensity for research [9]. In a more recent study on mathematical competitions, authors state that, through preparation and participation, students develop problem-solving competencies, improve their use of mathematical notation, and enhance their argumentation and proof-writing skills, as well as their ability to visualize mathematical concepts [6].

Mathematical competitions, differences in students' achievements based on gender, school conditions, and the type of classes they attend have also been analyzed by authors from Republic of Serbia [11, 23, 24, 25]. In a study conducted by Vulović, Milenković and Milikić [25] it was shown that 7th and 8th-grade students attending specialized mathematics classes in grammar schools achieved statistically significantly better results in national mathematics competitions.

AI tools, mathematics problems and mathematical competitions

Even before the use of artificial intelligence (AI) tools for solving mathematical competition problems, various technologies were used to tackle problems from national mathematics competitions. For instance, Ariño-Morera et al. [1] explored solving problems from the Austrian National Mathematics Competition for high school students using GeoGebra Discovery, identifying its advantages, limitations, and possibilities for further improving its implementation for competitive students.

With the development of both widely accessible and specialized AI tools, researchers have begun exploring their capabilities in solving mathematical problems. Previous research suggests that in some areas, such as mathematics, AI tools cannot be used with absolute certainty. An analysis of AI-generated solutions for a set of mathematical problems has revealed the limitations of the widely accepted tool GPT-4 [13], in solving complex mathematical problems [7]. Notably, one study concluded that an average student requires teacher assistance to master quadratic equations at the school level, as independent use of AI tools may lead to misconceptions [5].

Another recent study [26] highlights that Generative AI can be used for proving mathematical statements but stresses that students must critically assess the solutions, determining which parts to accept and which to question. In other words, they should not assume that every AI-generated proof is flawlessly correct. According to these authors, activities that students need to engage in when analyzing AI-generated proofs include [26]:

- “Conjecturing, drafting, and writing final proofs as fundamental skills in the process of constructing proofs,
- Evaluating GenAI’s responses to develop skills in proof comprehension and evaluation,
- Determining what to take from the provided information, which relates to taking autonomous actions, which is a critical skill instead of passive acceptance.”

Jia et al. [8] examined the level of support AI tools provide for mathematical problem solving. Specifically, for a set of algebra and geometry problems, they generated six suggestions and guidelines from a Chain of Thought (CoT)-enabled AI tool and six from mathematics teachers. These suggestions were then evaluated by experts, who generally rated the teachers’ responses higher. However, evaluators often could not definitively distinguish between the AI-generated and teacher-provided suggestions. Researchers also analyzed how the way tasks are presented to AI tools influences their performance, noting certain limitations when solving problems that include images as part of the formulation [18].

Torres-Peña et al. [21] demonstrated how AI tools such as ChatGPT, MathGPT, Gemini, and Wolfram Alpha can be used to improve students’ understanding of key mathematical concepts, such as calculus. When it comes to applying AI tools for solving problems from renowned mathematics competitions, researchers [4, 22] focused on solving problems from the International Mathematical Olympiad (IMO) using AI. Trinh et al. [22] designed the AI system AlphaGeometry to solve complex geometry problems without requiring human-annotated training data. This system synthesized millions of theorems, along with their proofs, to achieve performance comparable to top IMO competitors. The solutions generated by AlphaGeometry follow a logically structured approach and are often accompanied by appropriate figures. Remarkably, the tool successfully solved two geometry problems from the IMO (2000 and 2015) that are widely considered the most challenging of the past 25 years [22]. Subsequently, the AlphaGeometry2 model was developed, significantly improving the percentage of correctly solved IMO geometry problems from the

past 25 years to 84% – an increase of nearly 30% [4]. This model would have won a silver medal at IMO 2024. Similarly, AlphaProof was designed to solve formal mathematical problems by proving or disproving statements. When tested on IMO competition problems, AlphaProof successfully solved one of the most challenging problems at IMO 2024 and achieved a score that would have secured a silver medal in competition with human participants [3].

The relevance of this topic is further underscored by a team of 15 researchers [10] who created a dataset containing 860,000 pairs of competition math problems and solutions. This dataset was developed to train new AI models on a large-scale mathematical dataset. Thanks to the NuminaMath dataset, the team won the first-ever Artificial Intelligence Mathematical Olympiad, held in 2024.

ChatGPT, a general-purpose tool developed by OpenAI, has been tested on problems from the American Mathematics Competitions (AMC). The results of this study [19] indicate that ChatGPT performed differently depending on the mathematical domain of the competition problems. It achieved the best results in number theory and computational problems but struggled with geometry and combinatorics. On average, ChatGPT correctly solved between 50% and 60% of the problems, often failing to generate correct solutions for more complex tasks.

In 2024, OpenAI introduced o1-preview [14], an advanced language model designed for complex reasoning and problem-solving. The o1-preview model employs a “think before answering” strategy by generating internal chains of thought before providing a final response. Since it cannot process problems that include images, it is ideal for solving text-based problems. Its ability to accept input in LaTeX syntax makes it particularly useful for handling problems with complex mathematical notation. This model ranked among the top 500 competitors in the qualifying rounds for the Mathematical Olympiad [14].

A notable shift in mathematical reasoning performance has emerged with the introduction of OpenAI’s newer models, o1 [15] and o3-mini [16], each released at different stages of the company’s model development. Compared to earlier models such as GPT-4-turbo, both demonstrate improved consistency in multi-step reasoning and enhanced accuracy on competition-style problems. For instance, o3-mini (high) achieved 83.6% accuracy on the AIME 2024 benchmark, while o1 reached 74.4%, placing it among the top 500 competitors in the U.S. [16]. These results indicate a substantial advancement in the models’ ability to tackle challenging mathematical tasks beyond standard educational benchmarks.

Methodology

Mathematics competitions for high school students in Serbia

The National Mathematics Competition is organized by the Mathematical Society of Serbia and the Ministry of Education, Science, and Technological Development of the Republic of Serbia. The competition syllabus includes topics for all grade levels and competition stages, covering not only regular curriculum topics but also additional material studied in advanced mathematics courses. The competition program is publicly available.

The National Mathematics Competition represents the third level of mathematics competitions for high school students. The first level is the municipal competition, from which the best students advance to the district competition, which serves as the second level. Students who achieve the highest scores at the district level qualify for the national competition. The fourth level is the Serbian Mathematical Olympiad (SMO).

Students compete within their respective grades (four grades in total) and in two categories. In Category A, competitors include students from the Mathematical Grammar School, specialized mathematics departments following the Mathematical Grammar School curriculum, as well as other high school students who wish to participate. There are also certain specific circumstances under which students may compete in Category A, as outlined in the Regulations on Competitions for High School Students, which is publicly available on the website of the Mathematical Society of Serbia. Category B consists of students from general grammar schools and other schools that do not follow the specialized mathematical curriculum. The National Mathematics Competition hosts up to 150 students in Category A and up to 240 students in Category B.

The top 32 students from Category A at the National Competition qualify for the SMO. From there, a selection is made for the Balkan Mathematical Olympiad (BMO), and later for the International Mathematical Olympiad (IMO). The Serbian team at IMO 2024 consisted of one 2nd-year student, one 3rd-year student, and four 4th-year students, all from Category A. These students won four silver medals, one bronze medal, and one honorable mention at the IMO.

The National Competition is held annually in different sessions for the two categories, under school conditions. Students in Category A solve four problems, while students in Category B solve five problems, with both groups having 240 minutes to complete their solutions. The maximum number of points that students in both categories can achieve is 100. The problems are designed and graded by the National Committee for Mathematics Competitions for High Schools, appointed by the Mathematical Society of Serbia.

Material and methods

OpenAI has recently released two advanced language models o1, introduced in September 2024, and o3-mini, launched in January 2025 that are designed to improve reasoning and problem-solving across a wide range of domains, including science, technology, engineering, and mathematics (STEM). Both models belong to the newest generation of AI systems, developed to go beyond text generation and support more structured, multi-step reasoning.

The o1 model is positioned as a high-intelligence reasoning system. It utilizes reinforcement learning and chain-of-thought prompting strategies to handle tasks that require deep analytical thinking and complex problem decomposition. Additionally, o1 is integrated into ChatGPT subscriptions, offering users enhanced reasoning capabilities within the ChatGPT interface.

The o3-mini model, while smaller in scale and optimized for efficiency, builds

on the capabilities of o1. It is engineered to deliver strong performance on technical tasks – including mathematics and programming - while maintaining faster response times and lower operational costs. Available through both the ChatGPT interface and OpenAI’s API, it is intended to serve as a practical and scalable tool for everyday use in STEM-related work.

These two models represent the latest stage in the evolution of OpenAI’s reasoning-oriented tools. Given their architecture and design goals, they present promising candidates for evaluation in mathematics problem solving, particularly in the context of high-school level competition problems. In this study, we test o1 and o3-mini on a selection of tasks from Serbia’s National Mathematics Competition, which were not part of their training data, to examine their ability to solve complex, multi-step problems in a high-stakes academic setting.

Aim of the study

The aim of this study is to evaluate the performance of OpenAI’s o1 and o3-mini models on high-school level competition problems from the 2024 National Mathematics Competition in the Republic of Serbia. These problems were not included in the models’ training data, making them suitable for assessing the models’ generalization and reasoning abilities in unfamiliar, complex tasks.

The objective of the research is to rank the solutions produced by the o1 and o3-mini tools alongside the solutions of students who participated in the National Mathematics Competition in 2024, and to determine the potential performance of these tools in comparison with the best high school mathematics students in Serbia.

In order to test the quality and accuracy of the solutions provided by the OpenAI tools o1 and o3-mini on competition-level mathematics problems, whose difficulty corresponds to the problems from the National Mathematics Competition in the Republic of Serbia, we evaluated the solutions generated by these two tools on the exact problems that students in Serbia solved at the national competition in 2024.

Considering that students in all four high school grades compete in two categories, based on the curriculum they follow in their respective departments, and that students in Category A solve 4 problems, while students in Category B solve 5 problems, we assigned to both the o1 and o3-mini tools a total of 16 problems for Category A and 20 problems for Category B. This amounts to a total of 36 problems assigned to each of the two tools. Considering that all problems were formulated without accompanying images and that some formulations used mathematical expressions to represent certain equations, inequalities, and concepts, all problem texts were provided to the o1 and o3-mini tools in LaTeX syntax, in the Serbian language, without additional instructions. After the solutions were generated by these tools, all 72 solutions were reviewed and evaluated by a member of the National Mathematics Competition Committee for high schools, who was actively involved in both the creation of the given problems and the evaluation of students’ solutions at the National Mathematics Competition held in 2024 in Republic of Serbia. The research was conducted during January, February, and March of 2025.

Results

Before evaluating the accuracy of the solutions, we analyzed the structure of the solutions produced by the o1 and o3-mini tools.

By analyzing the nature of the solutions provided by o1, we can observe that each individual aspect of the solution is clearly separated, with a subheading, so that the person following the solution (whether a student or a teacher) can more easily follow and better understand the components of the solution. It is particularly important to emphasize that each solution includes a concise and unambiguous conclusion, which o1 reaches through intermediate steps in which it proves certain properties, equalities, divisibility, etc., and carries out the computational procedure. This conclusion is specifically highlighted at the end of the process.

The solutions produced by o3-mini are also detailed. In certain tasks, as for o1 tool, it is necessary to wait for the tool to begin generating the solution, during which it indicates how much time it spent reasoning. In this part of the solution, the tool essentially summarizes the given information and what is being asked, after which it often writes down the idea behind the solution, then carries out the procedure, and finally provides a clear conclusion, that is, the final answer to the problem.

Reasoned for 25 seconds ^

Breaking down equations
OK, let me see. I'm working through a system of three equations with x, y, and z in terms of a, b, and $M = |a+2024|$, aiming to solve for x, y, and z.

Subtracting equations
Mapping out a standard method by subtracting equation (1) from (2), aiming to isolate y and z, but a mistake seems to stall the progress.

Refining the equation
OK, let's simplify the left-hand side of the equation by removing x. This results in $(M-1)y$, which should match the term on the right-hand side, a.

Deriving expressions
I'm thinking through $y = a/(M-1)$. For special cases where $M=1$, a is either -2023 or -2025. Simplifying the equation to solve for z, I find $z = (b-a)/(M-1)$.

Figuring out x
OK, let's see. Plugging in y and z, the expression for x simplifies to $[a(M-1) - b]/(M-1)$, valid when M isn't 1.

Examining contradictions
I've realized that setting $a=-2023$ leads to a contradiction in the system, as the equations can't all align. This means there might be no solution in this specific sub-case.

Unveiling contradictions
OK, I've noticed that setting $a=-2023$ or -2025 leads to contradictions where the equations can't align. This indicates that there is no solution for either sub-case when $M=1$.

Figure 1.1. Detailed and accurate o1 tool solution of Problem 1 (part 1)

Figures 1.1–1.4 show the detailed solution process given by o1 tool for the following problem.

Problem 1 (intended for third-grade students competing in Category B)

Solve the system of equations depending on the real parameters a and b .

$$\begin{aligned}x + y + z &= a \\x + |a + 2024|y + z &= 2a \\x + y + |a + 2024|z &= b.\end{aligned}$$

Mapping out the solution

Interestingly enough, if a is neither -2023 nor -2025 , a unique solution for x , y , and z is found, with M identified as $|a+2024|$ and M not equal to 1 .

Sketching the solution

I've worked out that for any a other than -2023 or -2025 , the solution involves distinct formulas for y , z , and x , with M being $|a+2024|$ and M not equaling 1 .

Taking a closer look

Interestingly enough, if a isn't -2023 or -2025 , a valid solution emerges with specific formulas for x , y , and z . But for $a = -2023$ or $a = -2025$, constraints rule out any solution.

Analyzing constraints

I'm piecing together the solution for cases where a is neither -2023 nor -2025 . For these values of a , the system results in no valid solutions.

Analiza i odgovor

Neka radi lakšeg zapisivanja uvedemo oznaku

$$M = |a + 2024|.$$

Sistem je:

$$\begin{cases}x + y + z = a, \\x + My + z = 2a, \\x + y + Mz = b.\end{cases}$$

Uočićemo da ključnu ulogu ima vrednost $M - 1$. 

Stoga je rešenje u ovom slučaju (kada $M \neq 1$):

$$x = \frac{a(M-1) - b}{M-1}, \quad y = \frac{a}{M-1}, \quad z = \frac{b-a}{M-1}, \quad \text{gde } M = |a + 2024|, \quad M \neq 1.$$

To važi za sve realne a , b osim kada $M = 1$.

Figure 1.2. Detailed and accurate o1 tool solution of Problem 1 (part 2)

The problem belongs to the field of Algebra. From the presented solution, which is comprehensive and thorough, it can be observed that the given system of equations is first analyzed, along with the constraints and conditions that could lead to a contradiction. A substitution related to one of the parameters is introduced, after which the solution of the system is determined for different values of the parameters. Finally, an explicit conclusion is provided, highlighting the final solution to the problem.

1. Slučaj $M \neq 1$ (odnosno $a \neq -2023$ i $a \neq -2025$)

Tada $M - 1 \neq 0$. Možemo „oduzimanjem“ jednačina da pronađemo y i z .

- Oduzmemo prvu jednačinu od druge:

$$(x + My + z) - (x + y + z) = 2a - a,$$

odakle se sve osim koeficijenta uz y poništava, pa dobijamo

$$(M - 1)y = a \implies y = \frac{a}{M - 1}.$$

- Oduzmemo prvu jednačinu od treće:

$$(x + y + Mz) - (x + y + z) = b - a,$$

što daje

$$(M - 1)z = b - a \implies z = \frac{b - a}{M - 1}.$$

Najzad, iz prve jednačine $x + y + z = a$ sledi

$$x = a - y - z = a - \frac{a}{M - 1} - \frac{b - a}{M - 1}.$$

Zajedničkim imeniocem $M - 1$ dobijamo

$$x = \frac{(a(M - 1)) - a - (b - a)}{M - 1} = \frac{a(M - 1) - [a + (b - a)]}{M - 1} = \frac{a(M - 1) - b}{M - 1}.$$

- Iz prve dobijamo $x + y + z = a$.
- Iz druge dobijamo $x + y + z = 2a$.
- Iz treće dobijamo $x + y + z = b$.

Da bi sve tri mogle istovremeno da važe, morali bismo imati $a = 2a = b$. To bi zahtevalo $a = 2a \implies a = 0$ i zatim $b = a = 0$. Ali ovde je a fiksno -2023 ili -2025 , pa nikako ne može biti 0 . Dakle \emph{nema rešenja}.

Drugim rečima, za $a \in \{-2023, -2025\}$ (tj. kada $|a + 2024| = 1$) \emph{ni jedno} b ne omogućava rešenje — sistem je protivrečan.

Figure 1.3. Detailed and accurate o1 tool solution of Problem 1 (part 3)

Figure 2 shows the solution generated by the o3-mini tool for the following problem

Problem 2 (intended for second-grade students competing in Category A)

Chairs are arranged around a round table and numbered sequentially in a clockwise direction as $1, 2, \dots, n$, where $n \in \mathbb{N}$, $n > 1$. Each of these chairs is occupied by a person facing the center of the table, who either always tells the truth or always lies. Each person made a statement about whether their left neighbor is lying or telling the truth. Let us construct a word of length n , consisting of n letters, such that for all $1 \leq i \leq n$, the i -th position in the word contains the letter L if the person sitting in chair i stated that their left neighbor is lying, and

2. Slučaj $M = 1$ (odnosno $|a + 2024| = 1$)

Postoje dva podslučaja:

- $a + 2024 = 1 \implies a = -2023$.
- $a + 2024 = -1 \implies a = -2025$.

U oba slučaja $M = 1$. Tada sistem postaje:

$$\begin{cases} x + y + z = a, \\ x + 1 \cdot y + z = 2a, \\ x + y + 1 \cdot z = b. \end{cases}$$

Drugim rečima, i druga i treća jednačina se svode na

$$x + y + z,$$

ali sa različitim desnim stranama ($2a$ i b).

3. Zaključak

- Ako $a \neq -2023$ i $a \neq -2025$ onda postoji **jedinstveno** rešenje

$$(x, y, z) = \left(\frac{a(M-1) - b}{M-1}, \frac{a}{M-1}, \frac{b-a}{M-1} \right), \quad M = |a + 2024|.$$

(Ovde je $M - 1 \neq 0$.)

- Ako $a = -2023$ (ili) $a = -2025$ tada $M = 1$ i sistem **nema** rešenja ni za jednu vrednost b .

Figure 1.4. Detailed and accurate o1 tool solution of Problem 1 (part 4)

the letter T if the person sitting in chair i stated that their left neighbor is telling the truth. (The left neighbor of the person sitting in chair n is the one sitting in chair 1). A word of length n composed of the letters L and T is called admissible if it can be obtained using the procedure described above. How many different admissible words exist?

Reasoned for 19 seconds

Alright, we have a setup with a round table and n chairs, numbered clockwise. Each person either tells the truth or lies, and each person's statement is about their left neighbor, either calling them a liar or a truth-teller. A string of length n is formed where each position shows whether a person said their left neighbor was a liar ("L") or a truth-teller ("T"). The challenge is interpreting the truth based on these statements in the context of truth-tellers and liars mingling around the table.

Okay, so a "prolazna" word is a string achieved from a configuration of truth-tellers and liars sitting at a round table. Each person makes a statement about their left neighbor, saying "L" if they claim them to be a liar and "T" if they claim them to be a truth-teller. The question is how many distinct "prolazna" words can we get, given all possible configurations. The key challenge here is understanding the relation between individuals' types (truth-teller or liar) and their statements about their neighbors.

Alright, so we have a situation where each person's statement is determined by whether their type matches their left neighbor's. The statement is "T" (truth-teller) when their type equals the left neighbor's, and "L" (liar) otherwise. The word formed by these statements depends on these transitions — specifically, the number of "L"s corresponds to type changes in the circle, and the number of "T"s corresponds to consecutive same types. A key observation is that the word must have an even number of "L"s, which seems to be a necessary condition. But is it sufficient?

Označimo za svakog čoveka koji sedi na stolici i (gđje su stolice numerisane $1, 2, \dots, n$ u smeru kazaljke na satu) njegovu „vrstu“ — neka je $x_i = 1$ ako je čovek istinitosnažan (govori istinu) i $x_i = 0$ ako je lažljiv. Budući da je svaki čovek okrenut ka sredini, njegov levi sused je onaj koji sedi neposredno pre njega u kružnom poretku — konkretno, osoba na stolici 1 ima levog suseda osobu sa oznakom n , a za $2 \leq i \leq n$ levi sused osobe na stolici i je osoba na stolici $i - 1$.

Prema uslovu zadatka, svaki čovek je izjavio da li njegov levi sused laže ili govori istinu. Pri tom važi da:

- Ako čovek govori istinu (tj. $x_i = 1$), onda on će reći „istina“ ukoliko je njegov levi sused istinitosnažan ($x_{i-1} = 1$), odnosno „laže“ ukoliko je njegov levi sused lažljiv ($x_{i-1} = 0$).
- Ako čovek laže (tj. $x_i = 0$), onda će izreći suprotan iskaz, dakle će reći „istina“ ako je njegov levi sused zapravo lažljiv ($x_{i-1} = 0$), a reći će „laže“ ako je njegov levi sused istinitosnažan ($x_{i-1} = 1$).

U oba slučaja, možemo zapisati da je izjava čoveka na stolici i (koju označavamo slovom w_i) određena relacijom:

$$w_i = \begin{cases} I, & \text{ako } x_i = x_{i-1}, \\ L, & \text{ako } x_i \neq x_{i-1}. \end{cases}$$

(Indeksiranje se vrši ciklično, tj. uzimamo $x_0 = x_n$.)

Dakle, reč $w = w_1 w_2 \dots w_n$ se sastoji od slova I (ako se tipovi ne menjaju) i L (ako se tipovi menjaju). Značajna opservacija je da, pošto se radi o kružnoj listi, broju puta kada se dogodi promena (tj. kada je $w_i = L$) mora biti paran — zato što ako se krećemo oko kruga i svaki put se menjamo, da bismo se "vratili" na početni tip mora biti par promena.

Dakle, „prolazna reč“ (to jest, reč dobijena ovim postupkom) može biti dobijena ako i samo ako sadrži paran broj slova L .

Sada, koliko ima reči dužine n od slova I i L sa parnim brojem L ?

- Ukupan broj reči dužine n je 2^n .
- Poznato je da se među svim 2^n binarnim nizovima jednako mnogo (to jest 2^{n-1}) nalazi onih sa parnim i onih sa neparnom paritetom (broj L — tj. „promena“).

Stoga je broj prolaznih reči (odnosno onih reči koje se mogu dobiti opisanim postupkom) jednak

$$2^{n-1}$$

Figure 2. Detailed and accurate o3-mini tool solution of Problem 2

The problem belongs to the field of Logic and Combinatorics, with a well-

known context where some individuals are telling the truth and others are lying. Solving this problem also requires knowledge of combinatorial principles.

Interestingly, both tools, in the part where they display details related to “understanding the problem” and creating a strategy for solving it, write that section in English, while the actual solution process is written in the language in which the problems were provided – that is, in Serbian. Additionally, by simply looking at the structure of the solutions, it is noticeable that there is an alternating pattern of mathematical notation and explanations provided in textual form, which corresponds to the format most commonly used in mathematical literature, such as collections of math problems with solutions.

In order to examine the performance of the o1 and o3-mini tools, we used the students’ results as reference values. In this way, we were able to evaluate the results produced, i.e., the solutions to the problems. Table 1 presents the students’ results (arithmetic means, standard deviations, and medians), as well as the number of points that the tools would have achieved in the given competition based on the solutions they generated. To obtain a more precise picture, we also calculated the Z-scores for the given values.

Category	Grade	Students' Results on the Entire Test			AI Tools' Results and Z-Scores in Comparison to Students' Results			
		Mean	SD	Median	o1	Z-score	o3-mini	Z-score
A	1	40.10	23.11	37.50	37.00	-0.13	37.00	-0.13
A	2	30.39	21.82	25.00	48.00	0.81	38.00	0.35
A	3	33.41	29.04	25.00	83.00	1.71	80.00	1.60
A	4	21.25	20.23	19.50	21.00	-0.01	43.00	1.08
B	1	55.19	13.66	58.00	73.00	1.30	73.00	1.30
B	2	45.13	19.66	45.00	82.00	1.88	82.00	1.88
B	3	42.08	11.06	41.00	50.00	0.72	52.00	0.90
B	4	38.00	17.63	37.00	56.00	1.02	80.00	2.38

Table 1. Students’ results at the National Mathematics Competition and the performance of the o1 and o3-mini tools

Let us first consider the results of the o1 tool. It is noticeable that in eight cases (four grades for each category), the number of points achieved by the o1 tool was lower than the average number of points achieved by students in the corresponding category in only two cases. This was the case for the first grade of Category A, as well as for the fourth grade of Category A (where the difference was only one quarter of a point out of a total of 100 points). The number of points achieved by the tool in the second grade of Category A was almost 18 points higher, while the best result in Category A was achieved in the third grade, where the score exceeded the average score of the third-grade students by almost 50 points, i.e., by 1.71 standard deviations.

Regarding the results of the o1 tool in solving problems in Category B, in all cases, the results exceeded both the average values and the medians of the scores achieved by the students. The best results of the o1 tool were recorded in the first two grades of high school (73 and 82 points), which are 1.30 and 1.88 standard deviations higher, respectively, than the scores achieved by the students of those grades. For the third and fourth grades, o1 achieved half of the total possible points, i.e., 50 and 56 points, which is 0.72 and slightly more than one standard deviation, respectively, higher than the arithmetic mean of the scores achieved by the corresponding group of competitors who participated in the national mathematics competition.

Regarding the results achieved by the o3-mini tool, in only one of the 8 cases did it perform worse than the average score achieved by students in the given category and grade, and this was again the case in the first grade of Category A. In the remaining 3 grades of Category A, it achieved a score higher than the arithmetic mean of the points earned by the students by 7.61, 46.59, and 21.75, or by 0.35, 1.60, and 1.08 standard deviations, respectively, for the second, third, and fourth grades.

When solving problems for Category B, in all four grades, the tool exceeded the average total score achieved by the students. Except for the third grade, where the Z-score value was below 1 (0.90), in the first and second grades of Category B, the score achieved by o3-mini was more than one standard deviation above the arithmetic mean (1.30 in the first and 1.88 in the second), while in the fourth grade, the score was almost two and a half (2.38) standard deviations higher than the number of points achieved by the students. It was precisely in the fourth grade of Category B that one of the two observed tools exceeded the students' scores the most.

Category	Grade	Number of competitors	o1 ranking	prize	o3-mini ranking	prize
A	1	50	28 th	No prize	28 th	No prize
A	2	41	10 th	3 rd prize	14 th	3 rd prize
A	3	34	5 th	2 nd prize	5 th	2 nd prize
A	4	28	14 th	3 rd prize	6 th	2 nd prize
B	1	42	5 th	2 nd prize	5 th	2 nd prize
B	2	67	2 nd	1 st prize	2 nd	1 st prize
B	3	64	12 th	2 nd prize	10 th	2 nd prize
B	4	64	9 th	2 nd prize	2 nd	1 st prize

Table 2. Placement of the o1 and o3-mini tools in comparison to the competitors who participated in the National Mathematics Competition

We were curious about how these two tools would be ranked if their results

were compared with those of the students who qualified for the national mathematics competition in 2024.

In Category A, the o1 tool would be ranked in the middle of the list in the first grade and would not win any prize. In the second and fourth grades, o1 would win the third prize, as it would be ranked 10th and 14th, respectively. In the second grade, this tool would win the second prize, as only four students would have better results than it. In the competition for students in Category B, o1 tool would be in the top 20% of students in each grade. In the first, third, and fourth grades, it would win the second prize, while in the second grade, it would be ranked first prize, as only one student would have better results than it.

The o3-mini tool would not win any award in Category A in the first grade, as it would be in the middle of the list again. In the second grade, it would be around the third of the list, winning the third prize. In the third and fourth grades, it would win the second prize, ranking fifth and sixth on the lists (in the top seventh and fifth of the students' works). In competition with students from Category B, o3-mini would win the second prize in the first and third grades, ranking 5th and 10th on the leaderboard, respectively. In the second and fourth grades of high school, o3-mini would win the first prize, as it would rank second on the leaderboard in both cases.

Discussion

Based on the results obtained, the o1 tool would have been awarded two bronze and one silver medal, while o3-mini would have won one bronze and two silver medals in the four grade levels of Category A, from which the top students qualify for the Balkan Mathematical Olympiad (BMO) and later for the International Mathematical Olympiad (IMO). Analogously, in Category B, o1 would have won two silver and two gold medals, while o3-mini would have won as many as three gold and one silver medal. Although both tools achieved relatively strong results, o3-mini outperformed o1 across some categories. Similarly, models designed to solve even more complex problems than those considered in this study such as those tackled by students at the highest levels of competition (IMO) have also achieved results that would earn them medals (mostly silver) when competing against students [3, 4, 22].

When analyzed at the category level, o1 achieved an average of 47.25% of the possible points in Category A, while o3-mini achieved 49.5%. In Category B, the o1 and o3-mini tools scored 65.25% and 71.25% of the possible points, respectively, which closely aligns with the results that OpenAI's tool (specifically ChatGPT) achieved in American Mathematics Competitions [19]. Given that the average score for both tools is lower in Category A than in Category B, we can conclude that o1 and o3-mini perform better on relatively easier problems compared to more difficult ones once again aligning with the findings of the same authors [19].

The results of this study match with previous findings [7] that OpenAI tools have certain limitations in terms of the accuracy of mathematical problem-solving.

Neither o1 nor o3-mini managed to fully solve all problems. Given that not all solutions are completely correct, it follows that students should critically evaluate the solutions produced by o1 and o3-mini, carefully considering which parts of the solutions can be accepted as correct and which cannot [26].

Conclusion

Through this research, we aimed to determine to what extent two OpenAI tools, o1 and o3-mini, are capable of successfully solving problems that high school students tackled at the National Mathematics Competition in Serbia in 2024. Although it is evident that neither tool managed to solve all the problems correctly, it is noteworthy that, in seven out of eight cases (across two categories and four grade levels), both tools would have won prizes if their results were evaluated alongside the students' results. This finding supports the idea that the o1 and o3-mini tools have the potential to serve as valuable support resources for students preparing for national mathematics competitions. Specifically, these tools could provide inspiration and strategic guidance on problems for which students might initially lack an idea or approach. Of course, caution is necessary, as not all solutions generated by these tools are entirely accurate. However, considering that students who qualify for the national competition have already demonstrated a high level of mathematical knowledge and reasoning skills, it can be assumed that such exceptional students would be able to recognize potential errors and would not blindly accept the solutions provided by these tools. Naturally, this approach would require guided work with students, ideally through workshops organized in collaboration with mathematics teachers.

This consideration also defines the direction of future research - examining the effects of preparation programs for mathematically gifted students, particularly competition participants, that combine traditional instruction with the use of one of these two AI tools.

Acknowledgements. This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia through the Agreement No. 451-03-65/2024-03/200122.

REFERENCES

- [1] B. Ariño-Morera, Z. Kovács, T. Recio, P. Tolmos, *Solving with GeoGebra Discovery an Austrian mathematics olympiad problem: Lessons learned (2024)*, arXiv:2401.11906, 101–109.
- [2] J. R. Campbell, S. Cho, A. X. Feng, *Academic competition. A bridge to more opportunities for the talented*, Roper Review, **33**, 1 (2010), 5-7.
- [3] D. Castelvechi, *DeepMind hits milestone in solving maths problems – AIs next grand challenge*, Nature, **632**, 8024 (2024), 236-237.
- [4] Y. Chervonyi, T. H. Trinh, M. Olšák, X. Yang, H. Nguyen, M. Menegali, . . . , T. Luong, *Gold-medalist performance in solving olympiad geometry with AlphaGeometry2*, arXiv: 2502.03544, (2025), 1–28.
- [5] W. Daher, F. Gierdien, *Use of language by generative AI tools in mathematical problem solving: The case of ChatGPT*, African Journal of Research in Mathematics, Science and Technology Education, **28**, 2 (2024), 1–14.

- [6] M. F. De Losada, P. J. Taylor, *Perspectives on mathematics competitions and their relationship with mathematics education*, ZDM Mathematics Education, **54**, 5 (2022), 941–959.
- [7] S. Frieder, L. Pinchetti, R. R. Griffiths, T. Salvatori, T. Lukasiewicz, P. Petersen, J. Berner, *Mathematical capabilities of Chatgpt*, Advances in Neural Information Processing Systems, **36**, 27699–27744.
- [8] J. Jia, T. Wang, Y. Zhang, G. Wang, *The comparison of general tips for mathematical problem solving generated by generative AI with those generated by human teachers*, Asia Pacific Journal of Education, **44** (1), 8-28.
- [9] P. Kenderov, *Competitions and mathematics education*, Proceedings of the International Congress of Mathematicians, **3** (2006), 1583-1598.
- [10] J. Li, E. Beeching, L. Tunstall, B. Lipkin, R. Soletskyi, S. Huang, . . . , S. Polu, *Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions*, Hugging Face repository, **13**, 9 (2024).
- [11] V. Mičić, Z. Kadelburg, B. Popović, *The Mathematical Society of Serbia – 60 years*, The Teaching of Mathematics, **11** (1) (2008), 1-19.
- [12] M. Omerović, S. Rešić, A. Palić, T. Bazdalić, *Role of additional activities and competition in the teaching of mathematics*, Human Research in Rehabilitation, **10** (1) (2020), 41–50.
- [13] OpenAI (2023), *GPT-4*, Available: <https://openai.com/index/gpt-4-research/> [Accessed 10 March 2025]
- [14] OpenAI (2024), *o1-preview*, Available: <https://openai.com/index/introducing-openai-o1-preview/> [Accessed 10 March 2025]
- [15] OpenAI (2024), *o1*, Available: <https://openai.com/index/openai-o1-system-card/> [Accessed 10 March 2025].
- [16] OpenAI (2025), *o3-mini*, Available: <https://openai.com/index/openai-o3-mini/> [Accessed 10 March 2025].
- [17] D. Ozdemir, *An Examination of students' views about an international math contest*, International Electronic Journal of Mathematics Education, **17** (2), (2022), 1–10.
- [18] M. Stanković, A. Milenković, M. Svičević, N. Vučićević, *Performance of an AI Tool in solving non-standard mathematics competition problems*, In: A. Spasić, D. Stojanović (Eds.), *Education and Artificial Intelligence (EDAI 2024): Book of abstracts*, Vranje, University of Niš, November 2930, 2024, (pp. 49),.
- [19] R. R. Teegavarapu, H. Sanghvi, *Analyzing the competitive mathematical problem-solving skills of ChatGPT*, In: *2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, (pp. 1-7), IEEE, 2023.
- [20] T. N. Thrasher, *The benefits of mathematics competitions*, Alabama Journal of Mathematics, **33** (2008), 59–63.
- [21] R. C. Torres-Peña, D. Peña-González, E. Chacuto-López, E. A. Ariza, D. Vergara, *Updating calculus teaching with AI: A classroom experience*, Education Sciences, **14**, 9 (2024), 1019.
- [22] T. H. Trinh, Y. Wu, Q. V. Le, H. He, T. Luong, *Solving olympiad geometry without human demonstrations*, Nature, **625**, 7995 (2024), 476-482.
- [23] N. Vulović, A. Mihajlović, J. Milinković, *Polne razlike na matematičkim takmičenjima u Republici Srbiji* [in Serbian], Inovacije u nastavi, **36**, 3 (2023), 119-135.
- [24] N. Vulović, A. Mihajlović, M. Milikić, *Mathematical competitions and gender differences in the achievements of third-grade elementary school students*, In: N. Malinović-Jovanović (Ed.), *Quality of Teaching and Learning (QTL) - Reality and Possibilities*, November 1718, 2023, Vranje (pp. 33-41). Vranje: Pedagogical Faculty University in Niš.
- [25] N. Vulović, A. Milenković, M. Milikić, *Razlike u postignućima učenika 7. i 8. razreda koji nastavu pohađaju u osnovnim školama i pri gimnazijama na državnim takmičenjima iz matematike* [in Serbian], Zbornik radova Pedagoškog fakulteta u Užicu, **26** (2024), 221-244.
- [26] H. Yoon, J. Hwang, K. Lee, K. H. Roh, O. N. Kwon, *Students' use of generative artificial intelligence for proving mathematical statements*, ZDM-Mathematics Education, **56**, 7 (2024), 1531-1551.

A.M.: University of Kragujevac, Faculty of Science, Kragujevac, Serbia

ORCID: 0000-0001-6699-8772

E-mail: aleksandar.milenkovic@pmf.kg.ac.rs

N.V.: University of Kragujevac, Faculty of Science, Kragujevac, Serbia

ORCID: 0000-0002-4903-7280

E-mail: nemanja.vucicevic@pmf.kg.ac.rs

M.S.: University of Kragujevac, Faculty of Science, Kragujevac, Serbia

ORCID: 0000-0003-2791-3849

E-mail: marina.svicevic@pmf.kg.ac.rs

Received: 04.04.2025

Accepted: 04.06.2025